Importance Sampling combiné avec les algorithmes MCMC dans le cas d'estimations répétées.

Dorota Gajda[1,2] & Chantal Guihenneuc-Jouyaux[1,2,3] & Judith Rousseau[4] & Kerrie Mengersen[5] & Darfiana Nur[6]

[1] *Inserm U780, IFR69, Paris XI, 16 avenue Paul Vaillant Couturier Villejuif, F-94807, France*

[2] *Inserm U754, IFR69, Paris XI, Villejuif, F-94807, France*

[3] *Université Paris Descartes, lab. MAP5, CNRS UMR 8145, 45 rue des Saints Peres, 75006 Paris, France*

[4] *Université Paris Dauphine, Paris, France*

[5] *QUT, Brisbone, Australia*

[6] *School of Mathematical and Physical Sciences The University of Newcastle, Callaghan, NSW 2308, Australia*

Mots-clés : MCMC, Importance Sampling, Modèle de Poisson

Résumé

L'objet de ce travail est de présenter l'Importance Sampling comme une méthode d'optimisation algorithmique dans le cas de l'étude empirique (basée sur des simulations) d'un estimateur dans le cadre d'une modélisation bayésienne. L'étude de simulations permet d'évaluer certaines propriétés statistiques d'un estimateur via des réplications d'événements aléatoires. Afin de caractériser les performances des estimateurs selon différentes situations et de contrôler les fluctuations aléatoires, ce type d'analyse nécessite de simuler sous différentes paramétrisations beaucoup de jeux de données, puis pour chaque jeu de données, d'estimer les paramètres ou des fonctions de ces paramètres. Le contexte de notre travail est un modèle paramétrique sous lequel les jeux de données ont été simulés pour certaines valeurs des paramètres. Concernant l'estimation dans le contexte Bayésien, des lois a priori sont été spécifiées sur les paramètres, ces lois a priori restent les mêmes quelques soient les jeux de données. La démarche bayésienne, comme abordée dans la vaste littérature (Cf. par exemple Robert (2007)), consiste à combiner l'information a priori des paramètres représentée par des lois a priori avec la source d'information provenant des données à travers la vraisemblance pour obtenir la loi a posteriori des paramètres conditionnelle aux données. Quand la loi a posteriori ou quand les moments de cette loi ne sont pas explicites, une approximation est obtenue par des algorithmes stochastiques basés sur les méthodes dites de Monte Carlo par Chaînes de Markov (MCMC) comme présentées par Hastings (1970) ou Geman et Geman (1984). Ces algorithmes permettent d'obtenir des réalisations Markoviennes de la loi a posteriori recherchée et, via la théorie ergodique, d'obtenir ainsi des estimations de ses moments.

D'un point de vue pratique, le recours aux algorithmes itératifs MCMC doit être fait pour chaque jeu de données simulé. L'utilisation répétée des algorithmes itératifs MCMC peut être très coûteuse en temps calcul.

L'objectif de notre travail est d'étudier et d'améliorer l'efficacité de l'utilisation dans les inférences bayésiennes d'une autre technique basée sur l'Importance Sampling (IS) dans le contexte présenté ci-dessus. Cette méthode nécessite le choix d'une fonction d'importance, choix souvent délicat à faire. Dans le cas particulier de l'étude de différents de jeux de données, notre idée consiste à utiliser l'algorithme MCMC pour un nombre limité de jeux de données présélectionnés et ainsi d'obtenir des réalisations de chacune des lois a posteriori correspondantes. Les estimations des paramètres sous les autres jeux de données seront alors faites via IS en ayant préalablement choisi une des lois a posteriori présélectionnées. La fonction d'importance est donc ici la loi a posteriori choisie. L'idée d'utilisation simultanée de l'IS a été déjà proposée entre autre par Geyer et Thompson (1992), Gelfand (1992) ou plus récemment par McVinish et al. (2008) néanmoins dans des contextes différents.

Dans notre étude, nous avons simulé un total de 100 jeux de données et présélectionné 10 jeux de données. Nous avons testé deux stratégies de choix de la fonction d'importance appelée ici loi a posteriori de " référence " pour le calcul de l'IS : une, qui tire au hasard une seule loi a posteriori de référence pour toutes les estimations (appelée "référence fixe"), et l'autre qui choisit pour chaque nouvelle estimation (et donc chaque nouveau jeu de données) une loi a posteriori de référence parmi les dix premières présélectionnées (appelée "référence choisie"). Pour la deuxième stratégie, nous proposons deux critères de choix : Le premier basé sur la minimisation de la norme L_1 de la différence entre deux lois a posteriori et le deuxième basé sur la minimisation de la variance de l'estimation MCMC. Pour éviter le choix arbitraire des dix lois a posteriori présélectionnées, une procédure supplémentaire de sélection automatique a été établie. Les deux stratégies ont été comparées avec les résultats obtenus via MCMC sur la base d'erreurs quadratiques moyennes.

Les méthodes évoquées ici ont été étudiées sur trois types de modèles poissonniens: le modèle de Poisson avec un paramètre, la régression de Poisson avec une covariable (deux paramètres qui sont l'ordonnée à l'origine et le coefficient associé à la covariable), et la régression de Poisson avec extravariabilité (les deux paramètres précédents et en plus, la variance résiduelle). 100 jeux de données ont été simulés pour chaque modèle et pour des valeurs fixées des paramètres. Différentes valeurs des paramètres ont été choisies afin d'étudier des cas avec plus ou moins de variabilité. Nous avons attribué à tous les paramètres des lois a priori peu informatives, les mêmes pour toutes les estimations. Ensuite, pour chaque cadre de simulations, les moments de la loi a posteriori des paramètres ont été estimés de deux manières différentes: classiquement via MCMC et via Importance Sampling combiné avec MCMC comme présenté précédemment. Des résultats analytiques ont été possibles seulement dans le cas du premier modèle ce qui a permis de comparer les approximations aux vraies valeurs.

Les résultats montrent que les estimations via Importance Sampling fluctuent autour des valeurs des paramètres fixées dans les simulations (résultat attendu car l'information apportée par les données est importante et les lois a priori sont peu informatives). Les estimations IS sont, en général, proches des estimations obtenues classiquement via MCMC. Concernant la comparaison des différentes stratégies proposées avec IS, les erreurs quadratiques calculées avec les lois a posteriori de référence choisies par les critères, ont toujours été plus petites que les erreurs quadratiques obtenues avec la stratégie "référence fixe" quand la loi a posteriori de la référence fixe était la "pire". La procédure de sélection automatique, utilisée exclusivement pour le modèle le plus complexe, a donné des résultats comparables à ceux obtenus précédemment. L'utilisation des critères de choix dans le cadre des modèles étudiés avec les paramétrisations testées, a permis d'éliminer les cas des plus mauvaises approximations via IS avec référence fixe.

Abstract

The Importance Sampling method is used in combination with MCMC in Bayesian simulation study. In the particular context of numerous simulated data sets, MCMC algorithms have to be called several times which may become computationally expensive. Since Importance Sampling requires the choice of an importance function, we propose to use MCMC on a preselected set of the simulated data and then to obtain Markovian realisations of each corresponding posterior distribution. The estimates for the other simulated data are computed via IS by having previously chosen one of the preselected posterior distributions. This chosen posterior distribution is then the importance function.

IS procedure is improved by choosing for each data set a different importance function among the preselected set of posterior distributions. For each Importance Sampling estimation, we propose two criteria to select the suitable posterior distribution. The first criterion is based on the L_1 norm of the difference between two posterior distributions and the second one results from the minimization of the variance of MCMC estimate. A supplementary automatic selection procedure is also proposed to avoid arbitrary choice of the preselected set of importance functions. As a result we obtain estimations by MCMC and by IS, comparison being based on quadratic errors. The featured methods are illustrated in simulations studies under three types of Poisson models: simple Poisson model and two Poisson regression models with or without extra Poisson variability. Different parameter settings are considered.

Bibliographie

[1] Gelfand, A.E., and Dey, D.K., and Chang, H. (1992) *Model determination using predictive distributions with implementation via sampling-based methods*, Bayesian Statistics, Oxford University Press, 4, 147–167.

[2] Geman, S., and Geman, D. (1984) *Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, 721–740.

[3] Geyer,C.J., and Thompson,E.A.(1992) *Constrained Monte Carlo Maximum Likelihood for Dependent Data (with discussion)*, Journal of the Royal Statistical Society. Series B (Methodological), 54, 3, 657–699.
[4] Gilks, W.R., and Richardson, S., and Spiegelhalter D.(1996) *Markov Chain Monte Carlo in Practice*, Chapman & Hall.
[5] Hastings, W. K. (1970) *Monte Carlo Sampling Methods Using Markov Chains and Their Applications*, Biometrika, 57,1, 97–109.
[6] McVinish, R., and Mengersen, K., and Nur, D.C., and Rousseau, J., and Guihenneuc-Jouyaux, C. (2008) *Use of Importance Sampling for Repeated MCMC (to appear)*,
[7] Robert, C.P. (2007) *The Bayesian Choice. From Decision-Theoretic Foundations to Computational Implementation (Springer Texts in Statistics)*, Springer Verlag, New York.
[8] Gajda, D., Guihenneuc-Jouyaux C., Rousseau, J., Mengersen, K., Nur, D. (2009) *Use in practice of Importance Sampling for repeated MCMC for Poisson models (submitted)*